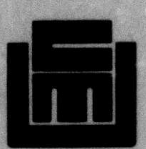
CMU-CS-82-140

Adding a zero-crossing count to spectral information in template-based speech recognition

Alexander I. Rudnicky, Alexander H. Waibel, and Neeraja Krishnan
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

DEPARTMENT
of
COMPUTER SCIENCE



Carnegie-Mellon University

CMU-CS-82-140

Adding a zero-crossing count to spectral information in template-based speech recognition

Alexander I. Rudnicky, Alexander H. Waibel, and Neeraja Krishnan
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

Zero-crossing data can provide important feature information about an utterance which is not available in a purely spectral representation. This report describes the incorporation of zero-crossing information into the spectral representation used in a template-matching system (CICADA). An analysis of zero-crossing data for an extensive (2880 utterance, 8 talker) alpha-digit data base is described. On the basis of this analysis, a zero-crossing algorithm is proposed. The algorithm was evaluated using a confusible subset of the alpha-digit vocabulary (the "E-set"). Inclusion of zero-crossing information in the representation leads to a 10-13% reduction in error rate, depending on the spectral representation.

This research was sponsored in part by the National Science Foundation, Grant MCS-7825824 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory Under Contract F33615-78-C-1551.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

1 Introduction[1]

An important consideration in the design of speech recognition systems is the choice of an accurate yet economical representation for the speech signal [Davis 80, White 76]. Most systems use a compact encoding of the short-term spectrum such as LPC or coefficients derived from band-pass filtering. Necessarily, a great deal of information (such as spectral detail and temporal structure) is lost in the encoding. For vocabularies containing words whose spectra are highly distinct (e.g., the digits) such a representation is adequate for high-accuracy recognition. In other cases, the coarseness of the spectral mapping leads to difficulties in discrimination. A subset of the alpha-digit vocabulary, words ending in the vowel /i/, illustrates such difficulties.[2] Utterances in the /i/ set are confusible because their distinctive characteristics are restricted almost entirely to a short segment at the beginning of the utterance (the consonant). It is in the nature of template matching to give equal weight to all portions of an utterance, as a result the contribution of the initial segment to the total distance between two utterances is frequently outweighted by random variations in the remainder of the utterance (the vowel). The goal of the work reported in this paper is to explore techniques that enhance the contribution of phonetically significant portions of an utterance while preserving a representation that allows a uniform template matching procedure to be used. The work described in this paper was done using the CICADA system developed at Carnegie-Mellon University [Alleva 81, Waibel 80]. CICADA uses a representation based on a compression of the short-term spectrum according to a 16 coefficient mel scale.

Let us consider the CICADA representation in more detail: In addition to a loss of fine spectral detail, two major features of the speech signal are lost in the mel-scale compression: The pitch of the vocalic portions and the distinction between a periodic and an aperiodic signal. Although the contribution of pitch information to phonetic identity is not well understood (see, however, [Massaro 78]), the latter distinction provides information that can be used to discriminate otherwise confusable utterances, such as "C"-"Z" and "T"-"D".

A purely spectral representation, particularly the kind used in the CICADA system, has only at best ambiguous information about the excitation source for a given speech segment. In this paper we describe an investigation of the potential advantages of including information about excitation source as a supplement to the spectral representation. We chose to investigate the zero-crossing

---

[1] *NOTE:* Some of the data described in this paper were reported earlier at the 102nd Meeting of the Acoustical Society of America, December, 1981

[2] The 10-member confusable set is composed of the letters "B","C","D","E","G","P","T","V","Z", and the digit "3".

count as a source of such information because of its straightforward derivation and its familiarity in the field (see e.g., [Baker 74]).

2 An analysis of zero-crossing statistics

The present section introduces the count method, describes a number of statistics for our speech corpus, and examines the use of statistical data as a basis for recognition decisions.

Zero-crossing statistics were collected for all utterances in our data base[3] (a total of 2880 utterances). The zero-crossing count was calculated using the same time-frame parameters used for the calculation of spectral coefficients; that is, over a 20 msec window stepped 10 msec through the utterance. The zero-crossing count was calculated using *non*-preemphasized speech. The potential range for the resulting zero-crossing count was thus 0-200, in actuality the observed range was 2-164. A number of alternate counting algorithms could have been used, most notably, calculating the zero-crossing count for only the central 10 msec of each spectral frame. Since the waveform is Hamming-windowed before spectral analysis, the central 10 msec would correspond to the region that makes the major contribution to the frame spectrum. On the other hand, a 10 msec window can give an unstable estimate of the zero-crossing count, thus introducing extra noise into the parameter. Empirical test seems to indicate that the 20 msec window results in better performance (see section 3).

Zero-crossing counts were collected for the utterance as bounded by the begin and end points determined by an automatic begin-end detector [Yegna 79]. The following statistics were calculated: the mean, the standard deviation, the median, and the range. In addition, the median zero-crossing count in ten equally-spaced intervals within an utterance was calculated.

Over the 8 talker database we observe a 2−164 range of zero-crossing counts. The highest mean and median zero-crossing counts are found for the utterance "SIX". This is to be expected, since "SIX" contains proportionately the most frication in the alpha-digit vocabulary. Looking at the standard deviations (SDs), we see again as expected, that utterances containing fricatives have high standard deviations. If we consider the absolute difference between the mean and the median to be a rough estimate of the skewness of a particular distribution, we note that pairs of utterances differing primarily in the degree of frication present (for example, C-Z, T-D, and P-B) show

---

[3] The database consists of 10 tokens of each word in the 36 member alpha-digit vocabulary (A..Z;0..9) recorded by 8 talkers (4 male, 4 female). All talkers were "naive". The material was recorded on audio tape in a moderately noisy ("office") environment then digitized using a 10 kHz sampling rate.

differences in the skewness of their zero-crossing count distributions: utterances with frication have more skewed distributions. Differences between utterances become even more apparent if we restrict the scope of our measurements to the initial portion of such utterances.

2.1 Using an utterance's zero-crossing count characteristics

The data described above suggest that we might be able to use the behaviour of the zero-crossing count in an utterance to supplement spectral matching. Unfortunately, these differences cannot be directly translated into reliable tests that distinguish pairs of spectrally similar utterances (for example, P-B). To show that this is the case, we will examine several specific instances.

Table 1 shows the *range* of standard deviations (SD) found for the minimal pairs B-P and D-T. If SD is to be used to reliably discriminate between the members of the two pairs, then the SD ranges for the two members of each pair must not overlap. As can be seen from the Table, this is the case only half the time. Several talkers (ds, gg, rp) show consistent discrimination for both pairs. For the other talkers, however, the separation is absent or is inconsistent across the two pairs examined. The same pattern is apparent for the skewness measure (Table 2): The members of a pair can be consistently distinguished for five talkers, inconsistently for two more and not at all for the remaining talkers. Focussing on those parts of the utterance known (from phonetic analysis) to carry the discriminative information, in this case the beginning of the utterance, does not allow us to realise an increase in accuracy or consistency (Table 3).

Table 1: The use of variance information for voicing discrimination

	utterance				decision	
dataset	B	D	P	T	P/B	T/D
ds	5–13	4–11	18–27	19–44	y	y
fa	4–11	5–12	5–18	8–22	n	n
gg	6–10	8–13	17–31	25–33	y	y
jl	6–14	6–18	8–26	29–39	n	y
ma	3–17	7–19	16–34	22–37	n	y
ms	3–11	8–28	4–21	16–32	n	n
rp	5–15	7–13	16–28	28–42	y	y
sw	3–19	7–18	9–21	11–22	n	n
Total correct classifications:					3	5

Note: values are the range of variances, calculated over 10 utterances for each talker.

This short exercise leads to the conclusion that zero-crossing information is not consistently useful

Table 2: Mean-median disparity

dataset	mean-median B	D	P	T	difference P/B	T/D
ds	3	2	9	14	y	y
fa	1	0	1	4	n	y
gg	1	2	11	12	y	y
jl	-2	-1	-4	-10	y	y
ma	2	4	9	14	y	y
ms	2	5	2	10	n	y
rp	0	0	9	17	y	y
sw	2	1	4	6	y	y
Total correct classifications:					6	8

Table 3: Zero crossing count range for voiced and unvoiced pairs

	voiced-unvoiced pairs ti	di	pi	bi	si	zi	non-overlaps
fa	34	56	8	32	58	140	0
ds	62	62	57	42	99	33	1.5
gg	75	37	52	31	35	67	2
jl	7	51	33	30	50	19	1
ma	55	57	37	35	70	102	1
ms	47	104	15	30	37	74	1
rp	38	46	84	59	30	81	1
sw	34	39	11	33	37	59	0
totals	1.5		5		1		

Note: The entires in this table indicate the relevant *extremes* of the range distribution. Thus, for the entries for the *unvoiced* members of each pair indicate the *lowest* zero-crossing count observed for that utterance, while the entries for the *voiced* member indicates the *highest* zero-crossing count for that utterance. Obviously, if the lowest unvoiced is less than the highest voiced count, zero crossing count is not diagnostic of voicing. Each value is based on ten (10) instances.

over all talkers or over all utterances. It is the case, however, that such information is consistent for some talkers and for some distinctions. Use of zero-crossing information, therefore, requires the presence of additional knowledge, such as might be obtained through tuning to a particular talker or through a recognition procedure that has the capability to narrow down choices to small sets for which the zero-crossing (or any other) attributes are well understood (see [Cole 81] for an example of this strategy). As an example of the latter, if the choices for an utterance can be narrowed down to "T-D", then the SD information can be used to pick either one word or the other. Likewise, if the choice can be narrowed down to "P-B", then the (raw) zero-crossing count for the initial portion of

the utterance can be used as a source of evidence. The implication is that given the imperfect reliability of zero-crossing information, it is best used in the context of a recognition strategy that incorporates detailed knowledge of the characteristics of speech sounds and uses an informed sequential decision strategy. While the C-MU speech group is actively pursuing work on such systems, the focus of the present paper is on the enhancement of template matching techniques.

One of the attractions of dynamic time warping is that it is a general decision procedure and makes no use of domain specific information. This allows the use of highly efficient search procedures, but requires that adequate discriminative information be included in the representations of the events being matched. The question therefore is whether a parameter such as the zero-crossing count can selectively enhance discriminative information contained in the representation.

Concretely, we are interested in determining the optimal manner in which to extract and transform zero-crossing information and include it as a part of a spectral template.

3 Designing an optimal zero-crossing function

The ideal zero-crossing coefficient would take on a neutral value when the speech signal was either silence or vocalic and go to one of several levels when different kinds of aperiodic energy were present in the speech. Such a function is in practice difficult to design, however, a reasonable approximation can be achieved. To do so, we must be able to define the following parameters:

- **a floor value:** a count below which it can be assumed no aperiodic energy of interest is present, i.e., during vocalic portions or during silence.
- **a ceiling value:** a count above which differentiating different degrees of zero-crossing counts is not informative.
- **quantization:** a mapping of the range between the ceiling and the floor which conveys useful information about the nature of the speech signal, e.g., providing distinctions between voiced and unvoiced fricatives.

The experiments described below represent a systematic, though certainly not exhaustive, search for optimal settings of the above parameters.

3.1 General procedure for all experiments

Except in the case of some pilot experiments (summarized in section 3.5), all experiments were performed using the full set of utterances (800) in the confusible "E" set. All matches were performed within talkers, using each of the ten repetitions of a vocabulary item in turn as a reference

template. Thus, a total of 900 matches were performed for each talker data set or a total of 7200 matches for each condition in the experiments described below. All experiments were performed using the CICADA2 system [Alleva 81].

3.2 The basic zero-crossing algorithm

A zero-crossing is defined as a transition between two successive waveform samples that produces a change in sign. The number of such transitions per frame is taken as the raw zero-crossing count for that frame. Except in the case of a few experiments (see section 3.5), the raw zero-crossing count is transformed by the following equation:

```
ZCcoefficient = (RawZC-Floor)/RangeFactor
```

The **Floor** and **RangeFactor** parameters correspond to the parameters described at the beginning of this section. Together they define a floor (explicitly) and a ceiling (implicitly) for the function. Values which fall outside this range are automatically clipped to the boundary values. In actual use, **ZCcoefficient** is normalized to a [-7..+7] range, corresponding to a 4 bit code. This feature is useful for our particular representation, as the spectral coefficients are also in this range. The contribution of **ZCcoefficient** to the calculation of a distance between two frames within the warping algorithm is equivalent to that of a single spectral coefficient, or one sixteenth. Differential weighting of **ZCcoefficient** is discussed in section 3.5.

3.3 Floor value

For our present purposes we will assume that the kind of information provided by the zero crossing count is rather limited in scope (but see [Baker 74]). That is, it can signal the presence of aperiodic speech energy; perhaps differentiate frication and aspiration, but no more. Zero-crossings, however, are always present in recorded speech: During periodic portions, during nominal silences, as well as during aperiodic portions. To make the count more sensitive, it is desirable to eliminate the "noise" introduced by zero crossings during silences and vocalic speech. One common technique is to calculate zero-crossing counts on the basis of a *center-clipped* signal. Although center-clipping has a number of advantages, we decided against using it because of its apparent sensitivity to changes in signal amplitude. In fact, one of our design considerations, being able to treat vocalic and silent segments equivalently, made the use of a "floor" threshold desirable.

A floor parameter is specified by selecting a threshold value for zero-crossing counts such that if the zero-crossing count falls below this value, it is assumed that no aperiodic speech energy is

present. To establish the level of "background zero-crossing" for voiced speech we can examine a vocalic utterance such as "ONE" and note the zero-crossing values encountered. A strict criterion would involve setting the threshold to the maximum zero-crossing value found in the all-voiced utterance. Although it would not be possible to guarantee that any value above this threshold indicates the presence of aperiodic energy (consider, for example, the interpolation of some environmental noise), the likelyhood of this being the case would be very high. Since a DC offset or the presence of voicing might alter the zero-crossing count for a segment that otherwise would be unambiguously identified as frication, it might be desirable to set a laxer criterion, preemphasize the signal, or even use the raw zero-crossing count.

The purpose of the present experiment is to establish an optimal value for this threshold, or `Floor` value. To simplify the design, the `RangeFactor` parameter was set to a constant value (4) that would ensure that most of the zero-crossing range was included, with minimal clipping at the ceiling value.

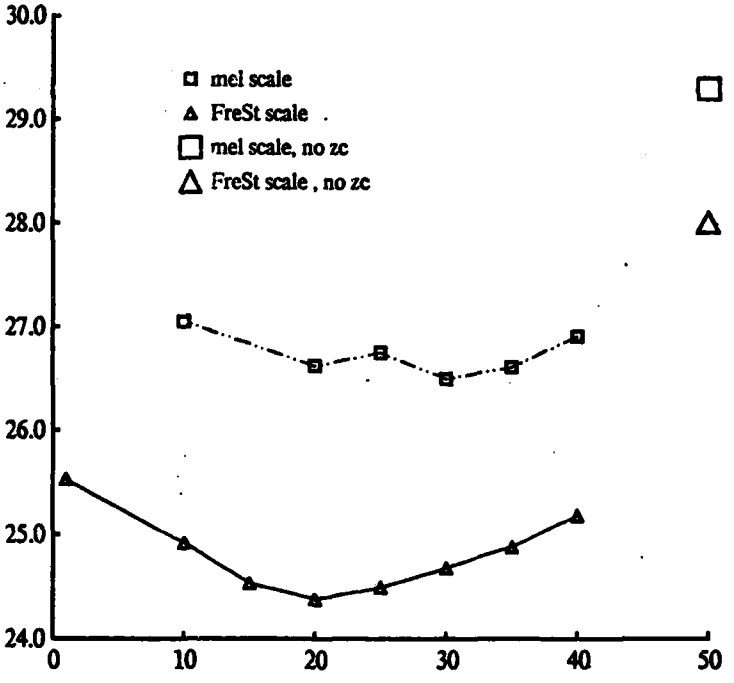


Figure 1: Performance (% error) as a function of floor value.

The experiment was performed using two spectral mapping scales: a 16 coefficient mel scale and a 16 coefficient scale based on the French and Steinberg [French 47] equi-intelligibility scale; this

latter scale will be referred to as the FreSt scale [Rudnicky 82]. The results are presented graphically in Figure 1. The optimal floor value for the mel scale is about 30, while the optimal value for the FreSt scale is about 20. The minima of the two functions are rather shallow and thus these values are approximate. There does not appear to be any ready explanation for the difference in minima for the two scales.

The present experiment establishes that the presence of a threshold (`Floor`) increases the accuracy of template matching. We believe that it does so by reducing the variance of the zero-crossing count in those parts of an utterance for which this information is not discriminative (i.e., vocalic segements). The reduction in variance contributes to a less-noisy match between the template and the test utterance.

3.4 The range factor

Zero-crossing counts have an inherent instability (as do all acoustic parameters of speech), it is therefore desirable to reduce their variability. This can be done by quantizing the range of the function and producing a smaller number of (discrete) levels. The present experiment compares several degrees of quantization by altering the `RangeFactor` parameter in the zero-crossing algorithm (see section 3.2). Table 4 shows the results of varying the range factor for two different floor values. Note that variations in the range factor appear to have a minimal effect on performance – all values obtained are within 0.5% of each other.

Table 4: Range Factor experiment Error Rates (%)

Range factor	Floor = 40	Floor = 20
3	--	24.83
4	25.18	24.38
5	25.24	24.38
6	25.33	24.57
7	25.69	--

Taken together with the `Floor` experiment, these data indicate that the full range of zero-crossing information (i.e., 2-164) is not necessary for effective use of this parameter. Thus, the (clipped) range between 20-65, divided into 15 levels, can give satisfactory performance. the interpretation of this result appears straight-forward: Low zero-crossing counts (i.e., below 20) are likely to come from the vowel portion of an utterance, since all vowels in the set studied are the

same, the zero-crossing fine-structure of the vowels does not contain useful information. (Although this may not be an appropriate conclusion, given that a constant vowel environment was used.) Similarly, zero-crossing counts of over 65 are almost certainly taken from aperiodic portions of an utterance, knowing an exact count above that ceiling does not provide any additional information and only contributes unnecessary variance. The range between the ceiling and floor values, however, may contain useful information about the degree of frication present in the signal. The results of the **RangeFactor** experiment suggests that, again, the fine-structure of the intermediate range contributes little specific information, the useful information being the fact that it is an intermediate range.

In order to understand the role of fine structure in the intermediate range, an additional experiment was performed, quantizing the intermediate range to successively coarser levels (from the original 15). Two levels were examined: 8 and 3. The mel scale spectral representation was used for this experiment. The **Floor** and **RangeFactor** parameters were set to 20 and 4, respectively. The results of this experiment are shown in Table 5.

Table 5: Range quantization

Number of levels	% error
3	28.77
8	25.55
15	25.22

As can be seen, the reduced number of levels leads to poorer performance. This result can be interpreted in one of two ways: Either the proportion of zero crossings present provides useful information and the coarse quantization destroys this information, or a gradual shift from one category to the other allows the recognition process to recover from errors of categorization. We believe that the latter is the case. An definitive assessment of this question, however, is beyond the scope of this paper.

3.5 Miscellaneous factors

This section describes several additional manipulations that were considered, but were not pursued. In all cases, the data were obtained for only three talkers (ds, fa, gg). These talkers were chosen to be representative of the entire 8 talker set. The results are displayed in Table 6. The 100 msec non-overlapped window condition uses an algorithm originally developed by [Niimi 80]. A number of findings are apparent: using the raw zero-crossing count degrades recognition performance, presumably because the inherent variability of the zero-crossing count adds more

noise than useful information. Doubling the weight also produces a decrement; this latter result is consistent with the results of the quatization experiment reported in section 3.4.

Table 6: Results of miscellaneous experiments

Manipulation	% error
mel scale only	27.81
mel scale; raw zero-crossing; 10msec non-overlapped window	30.41
mel scale; raw zero-crossing; 20msec overlapped window; value divided by 2.	31.18
FreSt scale only	28.01
FreSt scale; raw zero-crossing: 20 msec overlapped window	39.56
FreSt scale; floor = 30; ranged to [-14..+14], instead of [-7..+7] (i.e., coefficient weight doubled)	27.40

4 Analysing the improvement

The previous section has established that zero-crossing information can produce an improvement in recognition accuracy. The purpose of the present section is to examine in more detail the process by which zero-crossing information produces an improvement in recognition scores. Two questions will be dealt with: In which part of the utterance is the new, discriminative information present? How does zero-crossing information interact with the warping process?

4.1 The locus of useful zero-crossing information

The addition of zero-crossing information was hypothesized to enhance the discriminability of fricative and voiced portions of utterances. The utterances in the "E-set" differ only in the initial portion of the utterance and so we should expect that the improvement in performance is due to extra information at the beginning of the utterance. It is also possible, however, that for some reason utterances will differ in zero-crossing count not only at the beginning, but throughout the utterance. If this is the case, then we would be dealing with a qualitatively different phonetic difference than the one we were originally trying to represent (the fricative/non-fricative distinction). To assure that the improvement in performance was due to better discriminability based on the initial portion of the utterance, we performed the following experiment: All utterances were divided in half and a

recognition run was done separately on the first and second halves. If we are dealing with a whole-utterance phenomenon, then both halves should show some improvement in performance once zero-crossing information is added. If the additional information is present only at the beginning of the utterance, then only the first half of the utterance should show the improvement For this experiment, the previously determined optimal settings for the zero-crossing were used, i.e., `Floor`: 20, `RangeFactor`: 40. The results are shown in Table 7.

Table 7: Error rates for whole, 1st half, and 2nd half utterances

	whole	1st half	2nd half
FreSt scale only	28.01	24.29	73.27
FreSt scale and z-c	23.38	21.75	73.66
improvement	4.63	2.54	−0.39

As can be seen, the results support the hypothesis: Recognition based on only the second half of the utterance is equally poor, with or without zero-crossing information. In contrast, zero-crossing information improves performance for recognitions based on only the first half of the utterance. Two other interesting points should be noted about the data: There is an overall *improvement* in performance when only the first half is used, second, the improvement on first halves (with or without zero-crossing information) is *less than* the improvement obtained by adding zero-crossing information to the whole utterance. The improvement found when using only the first half of an utterance can be attributed to the elimination of mismatches caused by spurious differences at the ends of utterances. We believe these differences arise from instabilities in the speech signal that occur when phonation ceases.[4] In a template matching system, the resulting difference between reference and test utterances cannot be distinguished from meaningful differences, such as those which occur at the beginning of the utterances.

If, for a given unknown utterance, class membership (e.g., in the E-set) can be reliably determined, then it might be possible to improve recognition by focussing the recognition matching on the most informative portions of the utterance (see, e.g., [Niimi 80], and [Bradshaw 82]). The second result of interest is the relatively smaller improvement for the first half when zero-crossing information is added points to the asymptotic nature of most of the improvements which are added (see [Rudnicky 82] for further discussion of this point).

[4] In human speech perception, the information necessary for identifying an utterance must have presumably been extracted before the end of the utterance has been reached: Instabilities of this kind do not appear to influence the percept produced.

4.2 The influence of zero-crossing information on matching behaviour

The matching process in isolated word recognition can usually be factored into two conceptually distinct aspects: 1) the calculation of an optimal time alignment between an unknown utterance and a reference (i.e., establishing a warping path), 2) the calculation of a global "distance" between the two utterances, typically computed from inter-frame distances along the warping path. In practice, however, these two factors are confounded since the same measure, inter-frame distance, is used for both the choice of a warping path and for the calculation of distances. In the context of the present study, it is of interest to find out whether the increase in recognition accuracy is due to the development of a better warping path or to the enhancement of the score calculated along an existing path, or to a combination of the two.

With this in mind, an experiment was designed to separate the path-formation and distance formation aspects of the zero-crossing coefficient. The following four conditions were used:

1. Zerocrossing coefficient used for both path formation and distance calculation. (This is identical to the configurations used for the experiments in section 3).
2. Zerocrossing coefficient used for neither path nor distance.
3. Zerocrossing information used in conjunction with spectral information to determine best path. Distance calculated along the path *without* using the zero-crossing coefficient.
4. Spectral information *alone* used to determine the best path. Distance calculated with *both* spectrum and zero-crossing coefficients.

The results of the experiment are presented in Table 8. It is quite apparent that zero-crossing information contributes *only* to the distance calculation component of the matching process and not to the path-establishing component. There is no interaction between the two.

Table 8: Error rates for the path-distance experiment

		Distance		
		no	yes	difference
path	no	36.45	33.61	2.84
	yes	36.36	33.25	3.11
	difference	0.09	0.36	

We can draw the following conclusions from this experiment. Time alignment based on spectral information alone produces what appears to be an optimal alignment between a test and reference

utterance, at least one which cannot be improved upon by the addition of zero-crossing information. This result suggests that a fruitful approach to template matching might be the use of spectrum-based warping to align two utterances, followed by the use of this alignment to perform more detailed spectrum and feature-based distance computation over corresponding time frames (along the warping path) of the unknown and reference utterances. It also suggests that a promising avenue of exploration might be to determine the minimal information needed to produce optimal alignment, thereby freeing a system's resources to allow them to be concentrated on post-alignment processing.

5 Summary

The experiments described in this paper have shown that zero-crossing information can be successfully used to augment a spectral representation in a template-matching speech recognition system. The reduction in error rate, over 8 talkers, is between 10-13% depending on the spectral mapping. Taking into account that the specific parameter values presented in this paper are in all likelyhood not generalizable beyond the present recognition system, the alpha-digit vocabulary, and the set of talkers, we can offer the following guidelines for the use of zero-crossing information:

- Effective use of zero crossing information requires the elimination of "noise". This can be done by defining an "active" range for the count using floor and ceiling values. Once defined, this range can be effective even if quantized to a small number of levels.

- A promising strategy for template based recognition might be to separate time alignment and distance score computation. Simple spectral matching can be used to perform time alignment, while computationally more expensive discriminative techniques (based on feature extraction and phonetic knowledge) can be used to calculate the distance score.

6 References

[Alleva 81] F. Alleva.
Cicada2 Users' Manual.
1981.
Carnegie-Mellon Speech Group: Internal memo.

[Baker 74] Baker, J.M.
A new time-domain analysis of human speech and other complex waveforms.
PhD thesis, Carnegie-Mellon University, 1974.

[Bradshaw 82] G.L. Bradshaw, R.A. Cole, and Z. Li.
A comparison of learning techniques in speech recognition.
Technical Report, Carnegie-Mellon University, 1982.
(also presented at ICASSP'82).

[Cole 81] R.A. Cole.
A feature-based speech recognition strategy.
1981.
Carnegie-Mellon Speech Group: unpublished paper.

[Davis 80] Davis, S.B. and Mermelstein, P.
Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.
IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-28(4):357-366, August, 1980.

[French 47] French, N.R. and J.C. Steinberg.
Factors governing the intelligibility of speech sounds.
Journal of the Acoustical Society of America 19:90-119, 1947.

[Massaro 78] Massaro, D.W. and Cohen, M.M.
The contribution of fundamental frequency and voice-onset time to the /zi/-/si/ distinction.
Journal of the Acoustical Society of America 60:704-717, 1978.

[Niimi 80] Niimi, Y.
Weighed template matching.
1980.
Carnegie-Mellon Speech Group: unpublished research.

[Rudnicky 82] Rudnicky, A.I.
Representing the speech signal in a template matching system.
Technical Report, Carnegie-Mellon University, 1982.

[Waibel 80] A.Waibel, B.Yegnanarayana.
Optimization of Nonlinear Time Warping Techniques in Isolated Word Recognition Systems.
Technical Report, Carnegie-Mellon University, 1980.

[White 76] White, G.M. and Neely, R.B.
Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming.
IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-24:183-188, April, 1976.

[Yegna 79] Yegnanarayana, B.
An automatic begin-end detection algorithm based on spectral values.
1979.
CMU Speech Group: Internal memo.